DOCUMENT RESUME

ED 384 644 TM 023 848

AUTHOR Gomez, Leo; And Others
TITLE Natural Assessment of Oral Language Growth of Limited English Proficient Students in Paired Reciprocal Learning.
PUB DATE 19 Apr 95
NOTE 21p.; Paper presented at the Annual Meeting of the American Educational Research Association (San Francisco, CA, April 18-22, 1995).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Classroom Techniques; *Educational Assessment; *Elementary School Students; English (Second Language); Grade 5; *Hispanic Americans; Intermediate Grades; Interrater Reliability; Language Proficiency; *Language Tests; Learning; Limited English Speaking; *Naturalistic Observation; Oral Language; *Reciprocal Teaching; Second Language Learning; Spanish Speaking
IDENTIFIERS *Paired Student Interaction; Performance Based Evaluation; Time Series Analysis

ABSTRACT

The primary purpose of this study was the naturalistic assessment of growth in oral English proficiency of Hispanic American limited English proficient (LEP) students in a paired reciprocal learning format. Subjects were fifth-grade students in a summer English-as-a-Second-Language class. Twelve students in a paired learning format and 12 in a nonpaired format were evaluated in 4 weekly language samples with the Naturalistic Assessment of Oral English Proficiency (NAOEP) instrument developed for the study. The instrument contained eight language items for assessment and used a coding sheet for other significant data. Time series repeated measures analysis was used to detect significant changes in language growth. The NAOEP instrument exhibited acceptable interrater reliability but generally low score stability week-to-week. The instrument was not successful in determining language growth, specifically for the paired group, which demonstrated few gains on any variable. The naturalistic oral language assessment of LEP students appears to require a substantial number of classroom observations and a versatile language assessment instrument. Three tables, three figures, and a chart present study findings. (Contains 27 references.) (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

Department of Curriculum and Instruction
School of Education
University of Texas Pan American

ED 384 644

Natural Assessment of Oral Language Growth of Limited English Proficient Students in Paired Reciprocal Learning

A Paper Presented at the American Educational Research Association (AERA) 1995 Annual Meeting

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☒ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

by

Leo Gomez, Ph. D
Bilingual/ESL Professor
University of Texas Pan American

Richard Gomez, Ph. D.
Bilingual/ESL Professor
Texas Tech University

Rafael Lara-Alecio, Ph. D.
Bilingual/ESL Professor
Texas A&M University

PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
LEO GOMEZ
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

San Francisco, California

April 19, 1995

TM023148

BEST COPY AVAILABLE

Natural Assessment of Oral Language Growth of Limited English Proficient Students in Paired Reciprocal Learning

Purpose/Significance

The primary purpose of this study was the naturalistic assessment of growth in oral English proficiency of Hispanic limited English proficient (LEP) students in a paired reciprocal learning (PRL) format. Accurately assessing oral language proficiency of LEP children is a challenging one. One problem is determining which linguistic structures to describe (Silverman, Noa, & Russell, 1978). Whereas some assessments still rely on specific sub-skills and form (e.g. vocabulary, syntax, pronunciation), pragmatic assessment is increasingly influential (Hendricks, Scholtz, Spurling, Johnson, & Vandenburg, 1980; Mullen, 1980). The functional use of language to maintain and manipulate social situations has steadily become the focus of an increasing number of researchers (Doughty & Pica, 1986; Day, 1986; Long, 1981; Hatch, 1978). The theoretical framework that supports this movement is underscored by Hatch (1978): "One learns how to do conversation, one learns how to interact verbally, and out of this interaction syntactic structures are developed."

Since researchers have focused on the functional use of language for LEP learners to assess oral language proficiency, studies that support group interaction would prove beneficial. Alvarado (1992) asserts that in today's classrooms, increasing numbers of communicative tasks are being performed by small groups and pairs. She also states that the types of activities in second language classrooms have shifted to a more functional language approach with an emphasis on input and classroom interaction.

Peer/Paired Reciprocal Learning and L2 (second language) Acquisition

Today, there appears to be a shift from teacher talk dominated classrooms to creating classroom environments that encourage interaction between peers, providing greater exposure to the L2. Doughty and Pica (1986) illustrate this shift in their claim that as a means of increasing ESL students' target language practice time, teachers are establishing classroom environments that maximize use of small groups and pairs. Alvarado (1992) suggests that peer interaction through some form of pairing allows children to produce and receive more language and negotiate meaning through inter-group or inter-pair explanations and examples. Slavin (1988) reports that researchers continue to affirm that children are effective teachers because of their ability to communicate with one another both verbally and non-verbally. There is consensus among second language researchers that exposure to peers who speak the second language is valuable and can improve both the speed and kind of language acquired by limited English speakers (Hatch, 1978; McLaughlin, 1980).

Methodology

This study took place in a summer 5th grade ESL program during the summer of 1993 in a large Texas school district. The total study sample taken from a total program population of 107 was 24

students, with 12 students studied in a paired learning format (treatment group) and 12 in a non-paired learning format (comparison group). Student language levels were assessed and placed evenly in all classes.

Design/Procedure

This study utilized a replicated single subject repeated measures design including treatment and comparison groups. The repeated measures was four weekly language samples. Oral language proficiency was the dependent measure in this study: repeated passive assessment of oral language in a natural environment. Weekly classroom observations were conducted through video taping and observer field notes were taken during each session. Growth of oral language proficiency of each individual subject was measured by time series comparisons over the four week period. A total of four, 30 minute video tape segments of each individual subject, along with observer field notes, in natural classroom interaction was collected throughout the four week term. These notes were then analyzed and coded based on the NAOEP instrument in each of the eight variables for each five minute segment. Oral language samples were observed in the four major categorical areas of oral language proficiency: (a) communicative effectiveness, (b) fluency, (c) participation, and (d) frequency.

Instrument Development

In developing the NAOEP instrument, current instruments available for the assessment of grammatical and pragmatic language in social and academic situations were utilized for input. Various functional language assessment instruments were utilized for the creation of the assessment tool used in this study; Naturalistic Assessment of Oral English Proficiency (NAOEP). The creation of the NAOEP was based on the development of a classroom observation instrument able to measure oral language growth over time with reliability and, more importantly, *stability*. Well documented instruments such as Social Interaction Coding (Rice, Sell, & Hadley, 1987); Environmental Communication Profile (Calvert & Murray, 1985); Bilingual Oral Language Development (Mattes & Omark, 1984); Spotting Language problems (Damico & Oller, 1985); and Student Oral Language Observation Matrix (SOLOM) were analyzed for their effectiveness in assessing functional language in as naturalistic a context as possible.

The final NAOEP instrument was finally adopted for the study after considerable discussion by raters and several classroom observations and video recordings. The final instrument contained a total of 8 language items available for assessment and utilized a coding sheet for acquiring other significant data pertinent to the study (see attached).

Data Analysis

Data was analyzed using a Time Series Repeated Measures Analysis in order to detect significant changes in language growth (oral language proficiency) and language use for different purposes over the four week period. The following is a list of analyses conducted:

1) *KAPPA* (K) categorical statistic (Cohen, 1965), for determining inter-rater *reliability* of NAOEP instrument.

2) Subject, group, and total sample slope analysis to determine *stability* of NAOEP instrument over the four week period.
3) Two-way ANOVA with repeated measures, with pictorial graphs, for treatment and comparison group analysis to determine growth differences over the four week period and from week to week.

Results

#1: Inter-Rater Reliability

The reliability of the NAOEP instrument was based on classroom observations, tape recordings and constant negotiation between raters on each individual language variable. The reliability sample included a total of four students and five observations conducted over a two week period with constant modification and clarification of the instrument before statistically testing using Kappa. Due to the problems in using percent agreement only, which does not account for chance agreement, the *kappa* (k) categorical agreement statistic (Cohen, 1965) was used. One of Kappa's strengths is its ability to *correct* for chance agreement. Table 1 illustrates the results of kappa (k) statistic for categorical agreement using the NAOEP.

The KAPPA column indicates the final categorical measures on each of the eight variables accounting for chance agreement, and is the most conservative of all measures for inter-rater reliability. KAPPA is the final measure of agreement beyond chance. Note that KAPPA results for variables C: Topic Development (.426), E: Hesitations (.827), and H: Gestures/Expressions (.814) are equal to the best possible score that could have been obtained (KappaMax), controlling for chance agreement.

Also, note that the language variables which did not look at a specific oral language skill: E: Hesitations, G: Isolation/Participation, and H: Gestures/Expressions obtained the highest KAPPA results; .827, .643, and .814, respectively. Finally, Kappa/Kappa Max is the ratio of the KAPPA result over the best possible score (Kappa Max) for the particular language variable. Note that the KAPPA column depicts good reliability for the following language variables: E: Hesitations (.827) and Gestures and Expressions (.814). Even though the remaining language variables were not quite as high, they are all acceptable results once chance agreement is taken into account. Note that only the C: Topic Development variable scored below the .500 mark.

Table 1

*Kappa Statistic for Inter-Rater Reliability Using the NAOEP Instrument

Dependent Variable	Observed Agreement	Chance Agreement	KAPPA	Kappa/ Kappa Max	KappaMax
A Understand-ability & Sensibility	.690	.339	.531	.718	.739
B Providing Information	.733	.409	.549	.830	.662
C Topic Development	.667	.420	.426	1.000	.426
D Code Switching	.750	.427	.564	.660	.855
E Hesitations	.900	.423	.827	1.000	.827
F Grammar & Usage	.759	.329	.640	.675	.949
G Isolation/ Participation	.767	.346	.643	.716	.898
H Gestures/ Expressions	.909	.512	.814	1.000	.814

Note: All results based on a four point rating scale

#2: Stability Over Time

Question number two which sought to determine the stability of the NAOEP instrument (dependent measure) over a four week period utilized a slope analysis of time series data on each language variable. Assessing the stability of time series data is important because of the need to place trust in the instrument for further analysis. Scores that are very different from one day or period to another indicate that the instrument is not very reliable and therefore would probably provide inaccurate assessments of individual students. The stability analysis looked at the slope of a line of best fit through all data points over time. The analysis determined whether there was significant bounce of data points from the line of best fit.

Group slope analysis

Table 2 shows the results of the group slope analysis separately for the paired and non-paired groups. The purpose of this analysis was to determine if score stability over a four week period differed by groups. Note that the results of the paired group were not significant for all variables except for Grammar & Usage which was significant at .05, meaning that the slope was significantly different from zero. The results of the non-paired group were significant for dependent variables Grammar & Usage,

Table 2

Group Slope Analysis for Paired and Non-Paired Learners Per Dependent Variable

Dependent Variable	Independent Variable	Slope Coeff.	Std. Err.	T-Value
Understandability & Sensibility	Paired	-.06	.08	.39 NS
	Non-Paired	.49	.07	3.39 **
Providing Information	Paired	-.14	.07	.79 NS
	Non-Paired	.14	.08	.81 NS
Topic Development	Paired	-.37	.15	1.85 NS
	Non-Paired	.34	.13	1.73 NS
Code Switching	Paired	.06	.12	.36 NS
	Non-Paired	.05	.07	.30 NS
Hesitations	Paired	-.15	.13	.88 NS
	Non-Paired	.08	.13	.47 NS
Grammar & Usage	Paired	-.39	.08	2.64 *
	Non-Paired	.32	.07	2.03 *
Isolation/ Participation	Paired	.2	.10	1.26 NS
	Non-Paired	.36	.11	2.32 *
Use of Spanish	Paired	-.14	.14	.89 NS
	Non-Paired	-.02	.11	.09 NS
Self-Initiated Utterance	Paired	-.03	.25	.21 NS
	Non-Paired	.27	.35	1.72 NS

NS = Not Significant @ $p \leq .05$ * $p \leq .05$ ** $p \leq .01$

Isolation/Participation at .05, and Understandability & Sensibility at .01, meaning that the slopes for these variables were significantly different from zero. In addition, the standard error results of dependent variables Understandability and Sensibility, Providing Information, Code Switching, and Grammar & Usage for both groups indicate low error which means greater stability among these variables even if some were not significantly different from zero.

Individual slope analysis

Table 3 shows the results of the individual subject slope analyses per dependent variable. Computing a slope for individual subjects to determine stability would prove useful because an assessment instrument such as this would most likely be used at an individual level. As stated earlier, the stability of the slopes are indicated by standard error which is the area around the slope. The mean, standard error, t-value and p-value for each dependent variable was obtained for the least, the average, and the most stable sets of scores. Note that the mean and median columns indicate that no dependent variable was significantly different from zero and therefore exhibited low stability. The SE of the slope is one standard deviation value

plus or minus within the true slope actually lies. For example, the standard error score of Understandability and Sensibility at .15 for a subject at the mean indicates that there was high variability, or bounce, from one score to the next. In this case, we are 68% sure that the true slope lies plus or minus this standard error value. However, the low column shows that the dependent variables Sensibility and Understandability, Hesitations, and Self-Initiated Utterance were significant at the .05 level and stable for at least one subject. Also, note that variables Topic Development, Grammar & Usage and Isolation/Participation were highly stable for at least one subject at the .01 level.

Table 3

*Individual Subject Slope Analysis Per Dependent Variable

Dependent Variable		Low	Mean	Median	High
Understandability	Std. Error	.02	.15	.13	.34
Sensitivity	T-Value	.18 *	1.56 NS	1.47 NS	4.59 NS
Providing	Std. Error	.03	.16	.12	.50
Information	T-Value	.09 NS	1.81 NS	1.15 NS	9.39 NS
Topic	Std. Error	.04	.27	.2	.82
Development	T-Value	.1 **	2.93 NS	1.38 NS	11.49 NS
Code	Std. Error	.02	.17	.17	.49
Switching	T-Value	.71 NS	1.85 NS	1.73 NS	4.69 NS
Hesitations	Std. Error	.02	.19	.17	.48
	T-Value	.13 *	1.84 NS	1.73 NS	7.02 NS
Grammar & Usage	Std. Error	.02	.18	.16	.48
	T-Value	.15 **	2.28 NS	1.13 NS	16.50 NS
Isolation &	Std. Error	.001	.19	.18	.53
Participation	T-Value	.004 **	15.5 NS	1.07 NS	269.62 NS
Use of Spanish	Std. Error	.05	.32	.32	.73
	T-Value	.07 NS	.93 NS	.7 NS	2.57 NS
Self-Initiated	Std. Error	.11	.68	.64	1.85
Utterance	T-Value	.02 *	1.01 NS	.36 NS	4.20 NS

NS = Not Significant @ $p \leq .05$ * $p \leq .05$ ** $p \leq .01$

#3: Two-way factorial ANOVA & Graphs

In conducting the ANOVA analysis, it is important to point out that the ability to detect differences in growth between these two groups depends upon reliable measurement. In the previous section, two different reliability measurements were discussed; inter-rater reliability and stability over time. Although the reliability between raters was determined as acceptable, some significant problems in obtaining stability were noted. Instability will create difficulty in detecting differences between the two groups in their growth.

A two-way factorial ANOVA was the main analysis conducted to compare the growth of the paired and non-paired groups on each dependent variable. Specifically, the analysis examined whether there was a statistically significant difference between the two groups in their progress or growth over time. Based on the ANOVA results, mean scores were graphed to show the interaction. Interaction graphs were plotted primarily to illustrate the growth of each group since ANOVA only shows significant differences but not a specific group's growth over time.

The main purpose of the analysis was to determine the proportion of variance attributed to each independent variable (Paired learners, Non-paired learners and Time) and the interaction between them for each of the nine dependent variables. Finally, the graphs that follow each table provide a pictorial of the growth of the two groups over the four week period. The results of the ANOVA will be discussed according to their significance or non significance, based on the P-value results of each independent variable analysis for each dependent variable.

The most interesting result of the *ANOVA analyses was the interaction between the independent variables*Groups* and *Time*. Time x Groups, which compared the growth of the two groups over time showed that only **three** dependent variables exhibited significant differences in growth between the two groups at $P \leq .01$: (1) *Understandability & Sensibility*, (2) *Topic Development*, and (3) *Grammar & Usage*. Pictorial graphs of growth over time between groups for these three dependent variables will follow. The paired group, or treatment group (TG), is symbolized with a triangle and the non-paired group, or comparison group (CG), is represented with an (x). All other dependent variables resulted in no significant differences between both groups in their growth over time.

Understandability and Sensibility: The paired group steadily decreased in providing utterances that were understandable and sensible from week 1 to week 3, but demonstrated growth from week 3 to week 4. The non-paired group made impressive gains in providing understandable utterances from week 1 to week 3, but showed a slight decline from week 3 to week 4. (See Figure 1).

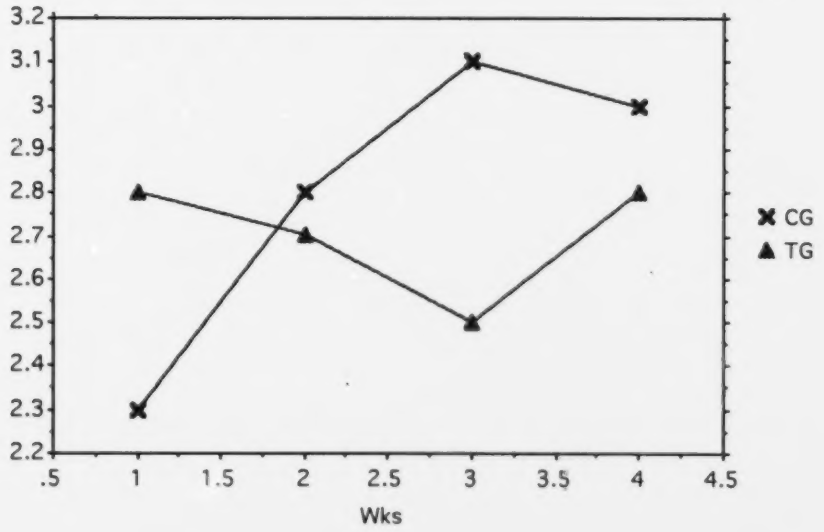
Topic Development: The paired group demonstrated significant growth in its ability to maintain a conversational topic over several exchanges from week 1 to week 2, but descended sharply from week 2 to week 4 ending up lower than original starting point. The non-paired group's original starting point for topic development was low, but gained steadily from week to week. (See Figure 2).

Grammar & Usage: The paired group sharply declined in correct grammar use from week 1 to week 2 and continued to slightly decline the remaining 2 weeks. The non-paired group slightly declined from week 1 to 2, but then showed steady growth in using correct grammar from week 2 to week 4 (see Figure 3)

Figure 1

Dependent Variable: *Understandability & Sensibility*

Independent Variables	df	Sum of Squares	Mean Squares	F-Value	P-Value
Time x Groups	3	2.10	.700	4.80	.006



Understandability & Sensibility

****

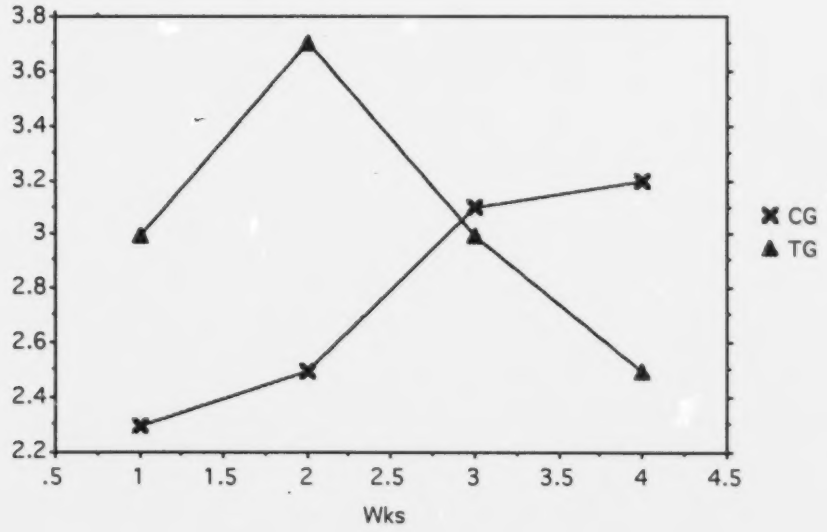
Paired Group: **Steadily decreased from wk 1 to wk 3, but demonstrated growth from wk 3 to wk 4**

Non-Paired Group: **Made impressive gains from wk 1 to wk 3, but slightly declined from wk 3 to wk 4**

Figure 2

Dependent Variable: ***Topic Development***

Independent Variables	df	Sum of Squares	Mean Squares	F-Value	P-Value
Time x Groups	3	1.600	.500	8.00	.0065



Topic Development

****

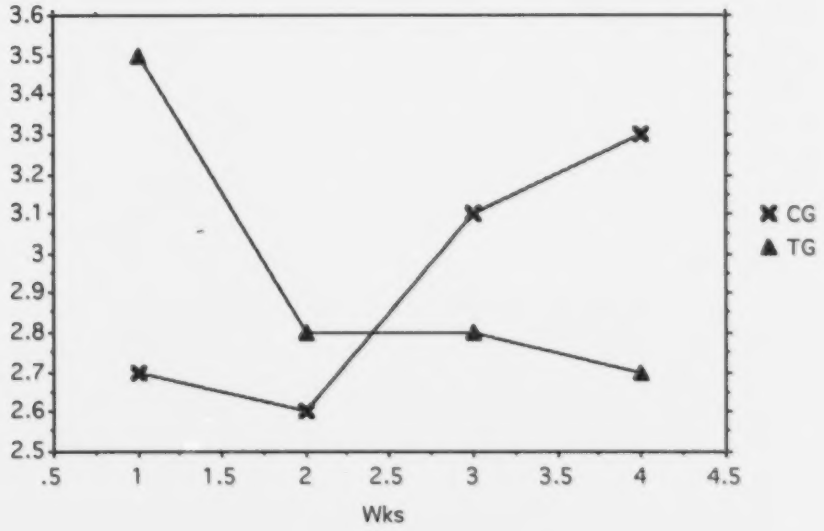
Paired Group: **Demonstrated significant growth from wk1 to wk 2, but descended sharply from wk 2 to wk 4, lower than original starting point**

Non-Paired Group: **Original starting point was low, but gained steadily from wk to wk**

Figure 3

Dependent Variable: *Grammar & Usage*

Independent Variables	df	Sum of Squares	Mean Squares	F-Value	P-Value
Time & Groups	3	4.200	1.400	7.83	.003



Grammar & Usage

****

Paired Group: **Sharply declined from wk 1 to wk 2 and continued to slightly decline**

Non-Paired Group: **Slightly declined from wk 1 to wk 2, but showed steady growth from wk 2 to wk 4**

Discussion

The primary concern of this study was to determine to what extent Hispanic limited English proficient students involved in a paired reciprocal learning format would improve their English oral proficiency as measured by the NAOEP instrument. In addition to this question, two preceding questions addressed inter-rater reliability and stability over time using the NAOEP instrument. The following is a discussion of the results of this study:

#1: Inter-Rater Reliability

The Kappa categorical agreement statistic was used to determine inter-rater reliability using the NAOEP instrument. Although final results indicated acceptable levels of inter-rater reliability for all language variables, particular variables deserve further discussion.

As displayed in Table 1, only two variables obtained observed agreement of 85% or higher: Hesitations (90%) and Gestures/Expressions (91%), followed by Isolation/ Participation (77%). This agreement may be explained by the common characteristics found among the three variables. Unlike the other language variables which assessed the accuracy of language use in one form or another, these three did not analyze oral language skills, but instead assessed physical actions or reactions by students. This type of assessment is judged with less difficulty since it is based on frequency and not on accuracy. The variable Topic Development, which obtained the lowest KAPPA score (.43), assessed the accuracy of conversational dialogue over several exchanges. This is a difficult task because of the speed in which conversations take place and/or use of gestures to convey meaning. This is supported by Higgs and Ray (1982) in their description of the use of language to convey meaning. They state that proper oral language assessment, for negotiation of meaning, becomes complicated when dealing with the accuracy of the linguistic effort.

This difference in inter-rater reliability between verbal and non-verbal variables suggests that more time be spent in assessing verbal interaction between L2 learners. In regard to the complicated task of assessing the accuracy of communication between peers, Tsui Bik-may (1987) states that the difficulty lies in the "well-defined assumptions which are shared between the speaker and the hearer." The observer, though in the same classroom, remains outside of the conversation. Therefore, proximity appears to be a major factor when assessing the accuracy of oral language use. It is important, as a rater, to be in a position to not only record the conversation verbatim, but also to be attentive to non-verbal communication cues.

#2: Stability Over Time

Group Slope Analyses

As explained earlier, both groups demonstrated a general lack of stability over the span of the study. However, the non-paired group showed better score stability for dependent variables: Code Switching, Use of Spanish, and Self-Initiated Utterances. This difference in stability may be explained by the characteristics inherent in the paired group which did not exist in the non-paired group.

One explanation for the greater lack of stability among the paired group may be the difference in language used for conversation vs. academic situations. Larger amounts of informal discussion may have occurred between paired subjects, for a variety of purposes other than academic, that did not occur among the non-paired subjects. Since non-paired subjects interacted more with the teacher than with a peer, similar use of academic language (focus on correctness) was more consistent from week to week for this group. Selinker (1972) explains that L2 learners tend to vary their developmental use of language to match the situation. He adds that language forms used in informal discourse are often very different from those used in formal discourse. Furthermore, this informal language use differs from one language situation to another. The Code Switching, Use of Spanish, and Self-Initiated Utterance variables support this informal use of language because these variables' attributes are inherently found in peer interaction within a naturalistic context.

The very nature of pairing and not pairing students may have had an effect on both the quantity and quality of interaction between the two groups. As explained earlier, one major attribute of the treatment group is the encouragement of student pairs to interact orally as much as possible. This interaction between pairs was often very subtle, including whispers, which often created difficulty in observer and video detection and/or interpretation. Moreover, approximately 80% of the interaction occurred within the pair and not to the teacher or aide. On the contrary, the comparison group primarily aimed their utterances at the teacher or aide, or at another peer not necessarily close to them. The results would be differences in the accuracy of language utterances due to observer proximity alone.

Individual Subject Slope Analysis

The individual slope analysis conducted for the mean indicated that no dependent variable was significantly different from zero and consequently exhibited low stability. Factors such as study length (4 weeks), final program week, number of measurements, and student absences possibly affected the results of this analysis. Although some individual subjects did exhibit high stability, the majority of students did not.

These results indicate that common factors affecting individual subjects may have also played a part in the stability of scores for most students. Ellis (1984 b), in his study on variability in interlanguage, states that individual L2 learners produce utterances which are formally different from one time to another, even when it is evident that they are performing the same communication function. He claims that this inconsistency greatly affects stable, accurate, and dependable language measurement.

Widdowson (1984) contends that the difficulty in obtaining stable language assessments from one time to another lies in the complexity inherent in language variations used by the learner for continued modification. He states that although linguistic competence has encompassed both appropriateness (communicative competence) and correctness of language use, instruments still fail to address the language learner's ability (which he calls *capacity*) to make his knowledge of linguistic rules work by using them in relation to both the situational and linguistic context. Capacity to construct

BEST COPY AVAILABLE

discourse, as defined by Widdowson, is an area which was not addressed in this study and requires further research.

Finally, another factor could be the effect the short amount of time had on the development of social relationships which was crucial for the encouragement of interaction. Both treatments in this study relied on consistent and natural classroom interaction in order to obtain reliable data. Unfortunately, lasting relationships were hampered by the shortness of time. This implies that ideally, effective oral language assessment for individual students should be conducted for a period of no less than 10 to 15 weeks.

#3: ANOVA Group Comparisons For Language Growth

In comparing the paired group and the non-paired group on language growth over time, two separate analyses were computed. The first of these analyses was a test for slope differences using the standardized coefficients of each dependent variable for each group. The second analysis was a two-way factorial ANOVA with repeated measures followed by group interaction graphs.

Two-Way Factorial ANOVA

ANOVA results indicate that there was indeed a significant difference in growth between the paired and non-paired group for variables measuring growth in verbal language use such as Understandability & Sensibility, Topic Development and Grammar & Usage. Although the treatment group (paired) initially scored higher on the scale during the first week for *all* three variables, gradual deterioration followed. On the other hand, the comparison group (non-paired) started out with very low scores in each variable but steadily improved from week to week. Possible explanations why the paired group did not improve in language growth as much as the non-paired group are: (a) incompatibility between pairs, (b) inexplicit conversations between pairs, (c) paired subjects taking greater language risks, (d) teacher differences, (e) learner differences, and (f) length of study.

Incompatibility Between Pairs

One explanation that may explain why the non-paired group improved better than the paired group was the personal differences found in the paired group. It was possible that many paired subjects were incompatible and thus unable to sustain lasting and/or meaningful conversations because of personality or motivational differences which in turn negatively impacted language growth. Another possibility is that paired subjects became bored with one another as time went on, causing a reduction in language interaction within the pair.

Inexplicit Conversations Between Pairs

A second explanation could be that being paired added the element of ambiguity in conversation. Paired subjects were not as clear and precise in their use of speech since they were communicating mainly with each other. Furthermore, there may have also been a tendency by the pair to speak softly and respond with short phrases that may have made sense only to the language partner. This type of interaction would have

produced inconsistencies in language use from week to week, resulting in erroneous measurement over the span of the study.

Paired Subjects Took Greater Language Risks

A third explanation why the non-paired group improved better than the paired group was the luxury that the paired subjects had over the non-paired subjects to take risks in their use of speech, because they were not being scrutinized by the teacher or another peer. Risk taking in second language acquisition, specifically in peer interaction, is consistent with what Rubin (1975) describes as *calculated guesses*, where the good L2 learner in peer negotiation makes willing guesses in order to communicate. Similarly, Beebe (1983) cites a study which claimed that moderate L2 risk takers have a greater motivation to achieve, but at times feel that it is safer to stay within patterns that communicate meaning even though there may be some errors in those patterns.

Teacher Differences

Teaching differences may have also contributed to the larger gains in language growth made by the non-paired group vs. the paired group. With an n of only 3 teachers for each group, it's possible that the treatment group (paired) teachers were the less skilled teachers in establishing and encouraging language interaction. Although fidelity of implementation was checked, no attempt was made to actively equalize teacher encouragement of language use. Johnson (1983), in her examination of the effects of Inter-ethnolinguistic Peer Tutoring (IEPT) on the social interaction and English proficiency of limited English speakers, concluded that social interaction may be influenced by the way educators structure classroom groups and activities.

Learner Differences

As language is acquired, the rate of acquisition from one L2 learner to another varies (Richard-Amato, 1988). Children acquire lexical and grammatical skills at a different pace. Although subjects were randomly assigned to both groups, this difference in acquisition may have also had an impact on the different language gains made by students in each group. The possibility exists that the paired group consisted of a number of subjects that acquired language at a slower pace. This difference may have been amplified by the pairing format itself. The fact that paired learners were instructed to work together, adds the element of cooperative interaction where the more proficient learner tends to dominate pair/group discussions (this can be verified by comparing data across pair members). This cooperative interaction limits the amount of speech produced by the less proficient. Non-paired subjects, however, were not limited or dependent on any one student(s), encouraging all student proficiency levels to verbalize.

Length of Study

Finally, the short length of time (4 weeks) may have also played a role in this analysis because more time would allow for more measurements, capturing a more realistic picture of the interactions that took place. Although this factor would have affected both groups, it may have affected the paired group more because time is essential to establish close relationships. Paired subjects may not have been prepared

to interact primarily with the same subject, and took time to adjust to this format. An example of this is depicted in the variable Understandability and Sensibility, which indicates that the paired group started out slow from week 1 to week 2, but improved from week 3 to week 4. Although we are not certain of the outcome, there is potential indication of growth for this variable.

Conclusion

The findings in this study indicate that the NAOEP instrument used to assess the use of oral language of LEP learners in natural context exhibited acceptable levels of inter-rater reliability, but generally low score stability from week to week for individual, group and total sample slope analyses. In addition, the instrument was unsuccessful in detecting significant growth over the span of the study for both groups, but specifically the paired group, which demonstrated few gains on any variable. Although the non-paired group appeared to demonstrate some growth in language over the four weeks, it was still not significant growth.

Of all results, lack of stability in scores throughout the study greatly affected the trustworthiness of the NAOEP instrument. Future research must be done on the development of low-inference descriptors that still address complicated and sophisticated language used in social and natural unstructured conversation, with some degree of stability.

The NAOEP instrument can be used by teachers for informal oral language assessment of individual students as they interact in pairs or as a group. It can be used primarily in classroom situations where the teacher works with a limited number of LEP children. However, the impracticality of the instrument, (e.g. equipment, coding time, etc.) limits its use with a large number of students.

It can be concluded from this study that naturalistic oral language assessment of LEP students requires a substantial number of classroom observations, and a versatile language assessment instrument in order to effectively examine the many variations of language used by peers, especially in paired reciprocal interaction. As suggested by Morrow (1985), effective oral language assessment must include long periods of observation in order to properly address the negotiation of language between L2 learners. Provided with more time, perhaps dependent variables related to the accuracy of oral language use such as Understandability & Sensibility, Topic Development and Grammar & Usage would demonstrate more growth for both the paired and non-paired groups.

Although the paired group demonstrated minimal gains in the English language over the four weeks, the greater interaction that developed from this format was evident. Paired learners shared a variety of language that may not have been produced had they not been in pairs. Perhaps adding one more subject to the pair would have created still greater interaction since the conversation is now promoted three ways.

This study supports the need to place careful thought to the current practices used in assessing language proficiency in second language acquisition programs and classrooms. As supported by the literature, and by the results on the difficulty in coding oral language within peer interaction, language

proficiency must be determined and assessed as it is used in natural context. A variety of language manifestations, which include both pragmatic and grammatical use should be the focus of the observation. Suggested future directions for research and instrument development include the distinction and integration of language descriptors which address the theoretical framework for communicative competence and peer interaction. Future language assessment instruments should focus on these two functional uses of language in natural context.

Caution should be taken with interpretation of findings due to several factors that have possibly influenced the results of this study. Specifically, lack of stability of scores, short amount of time, number and length of weekly observations, and lack of sensitivity of the NAOEP to effectively measure language growth must all caution generalizability of these results.

NATURALISTIC ASSESSMENT OF ORAL ENGLISH PROFICIENCY (NAOEP)

A. Understandability & Sensibility of Utterance.

1	2	3	4
Utterance is nonsensical and/or incomplete thought.	Utterance is sensible but lacking complete thought.	Utterance is sensible with a clear complete thought.	Utterance is sensible and advances the discussion. As a native English speaker.

*Always code this item, unless nothing was spoken at all.

B. Providing Information.

1	2	3	4
Cannot provide sufficient accurate information requested by a listener.	Provides some information required or requested by a listener.	Satisfactorily provides information required or requested by a listener.	Provides information beyond that required by a listener. As a native English speaker.

*Code only if subject responds to a request by another person.

C. Topic Development.

1	2	3	4
Cannot develop a topic beyond a few utterances. Erratically halts or changes topic.	Can develop a topic beyond a few utterances, but with some delays or misunderstandings.	Develops a topic over several exchanges, but cannot lead in topic development.	Smoothly and effortlessly leads in topic development. As a native English speaker.

*Code only if dialogue consists of at least 2 responses to person or group.

D. Code Switching.

1	2	3	4
Switches from one language to another within a given utterance, provided lengthy utterances took place.	Occassionally switches from one language to another within a given utterance, provided lengthy utterances took place.	Rarely switches from one language to another within a given utterance, provided lengthy utterances took place.	No code switching from one language to another, provided lengthy utterances took place. As a native English speaker.

* Always code this item.

E. Hesitations.

1	2	3	4
Frequent long and awkward delays before attempting to respond or during utterances.	Occasional delays (short and longer) before or during utterances.	Few occasions of short hesitations before or during utterances.	No hesitations before or during utterances. As a native English speaker.

*Always code this item. (Only hesitations based on oral responses/communications)

F. Grammar & Usage.

1	2	3	4
Frequent errors in basic grammar and usage (tense, gender) or major errors in vocabulary (word meaning).	Occasional errors in basic grammar and usage (tense, gender) or major errors in vocabulary (word meaning).	Practically no errors in basic or more advanced grammar, usage or vocabulary.	No errors in basic or more advanced grammar, usage or vocabulary. As a native English speaker.

*Always code this item. (50%+=1, 25%-50%=2, 25%-=3)

G. Isolation/Participation.

1	2	3	4
Isolates self—total avoidance of discussion/verbal exchanges.	Rarely engages in verbal exchanges. Hesitates to participate.	Occasionally engages in discussion/verbal exchanges without hesitation.	Consistently engages in discussion/ verbal exchanges without hesitation.

*Always code this item.

H. Gestures, Body Language, Humor, Expressions.

1	2	3
Uses gestures, body language, humor, and expressions rather than oral communication. (2 or more)	Some use of gestures, body language, humor, and expressions rather than oral communication. (1 time)	Gestures, body language, humor, and expressions are usually appropriate rather than oral communication. As a native English speaker.

*Code this item only when responding to verbal information without the use of oral communication.

BEST COPY AVAILABLE

Selected References

Alvarado, C. S. (1992). Discourse styles and patterns of participation on ESL interactive tasks. *TESOL Quarterly* , 26, 589-593.

Cohen, J. (1965). Some statistical issues in psychological research. In B. B. Wolman (Ed.), *Handbook of Clinical Psychology* (pp. 95-121). New York: McGraw-Hill.

Damico, J. S. (1991). Performance assessment of language minority students. *Proceedings of the Second National Research Symposium on Limited English Proficient Student Issues: Focus on Evaluation and Measurement*. Washington, D. C.: U. S. Dept. of Education (OBLEMA), 137-171.

Day, R. R. (1986). Introduction. In R. R. Day (Ed.), *Talking To Learn: Conversation in Second Language Acquisition* (pp. 3-4). Rowley, Mass: Newbury House.

Doughty, C., & Pica, T. (1986). "Information gap" tasks: Do they facilitate second language acquisition? *TESOL Quarterly*, 20, 305-325.

Ellis, R. (1984b). Sources of variability in interlanguage. Paper presented at the *Interlanguage Seminar in Honor of Pit Corder*, Edinburgh.

Hatch, E. (1978). Discourse analysis and second language acquisition. In E. Hatch (Ed.), *Second Language Acquisition: A Book of Readings* (pp. 401-435). Rowley, Mass.: Newbury House.

Hatch, E. (1992). *Discourse and Language Education*. New York: Cambridge University Press.

Hendricks, D., Scholz, G., Spurling, R., Johnson, M., & Vandenburg, L. (1980). Oral proficiency testing in an intensive English language program. In J. Oller & K. Perkins (Eds.), *Research in Language Testing* (pp. 77-90). Rowley, Mass.: Newbury House.

Higgs, T., & Ray, C. (1982). The push toward communication. In T. Higgs (Ed.), *Curriculum, Competence, and the Foreign Language Teacher* (pp. 57-79). Lincolnwood, IL: National Textbook Company.

Johnson, D. M. (1983). Natural language learning by design: Aclassroom experiment in social interaction and second language acquisition. *TESOL Quarterly*, 17, 55-68.

Krashen, S. D. (1985). *Inquiries and insights: Second language learning*. London: Longman.

Krashen, S. D. (1985a). *The input hypothesis: Issues and implications*. London: Longman.

Long, M. H. (1981). Input, interaction, and second language acquisition. In H. Winitz (Ed.), *Native Language and Foreign Language Acquisition* (pp. 259-278), Annals of the New York Academy of Sciences.

Long, M., & Porter, P. (1985). Group work, interlanguage talk, and second language acquisition. *TESOL Quarterly*, 19, 207-227.

McLaughlin, B. (1987). *Theories of second-language learning*. London: Arnold.

Morrow, D. G. (1985). Prominent characters and events organize narrative understanding, *Journal of Memory and Language*, 24: 304-319.

Mullen, K. A. (1980). Rater reliability and oral proficiency evaluations. In J. Oller & K. Perkins (Eds.), *Research in Language Testing* (91-101). Rowley, Mass.: Newbury House.

Oller, J. (1991). Language testing research: Lessons applied to LEP students and programs. *Proceedings of the Second National Research Symposium on Limited English Proficient Student Issues: Focus on Evaluation and Measurement*. Washington, D. C.: U. S. Dept. of Education (OBLEMA).

Pica, T., & Doughty, C. (1988). Variations in classroom interaction as a function of participation pattern and task. In J. Fine (Ed.), *Second Language Discourse: A Textbook of Current Research*, (pp. 41-55). Norwood, NJ: Ablex Publishing Corporation.

Pica, T., Young, R., & Doughty, C. (1994). The impact of interaction on comprehension. In R. M. Barasch & C. V. James (Eds.), *Beyond the Monitor Model* (pp. 97-119). Boston, MA: Heinle & Heinle Publishers.

Richard-Amato, P. (1988). *Making it happen: Interactions in a second language classroom*. New York: Longman.

Rubin, D., & Mead, N. (1984). *Large scale assessment of oral communication skills*. Urbana, IL: ERIC Clearinghouse on Reading and Communication Skills.

Selinker, L. (1972). Interlanguage. *International Review of Applied Linguistics*. 10, 209-230.

Slavin, R. (1988). Synthesis of research on grouping in elementary and secondary schools. *Educational Leadership*, 46, 67-77.

Tavener, J., Glynn, T. (1989). Peer tutoring as a context for children learning English as a second language. *Language and Education*, 3(1), 45-55.

Widdowson, H. (1984). *Learning Purpose And Language Use*. Oxford: Oxford University.